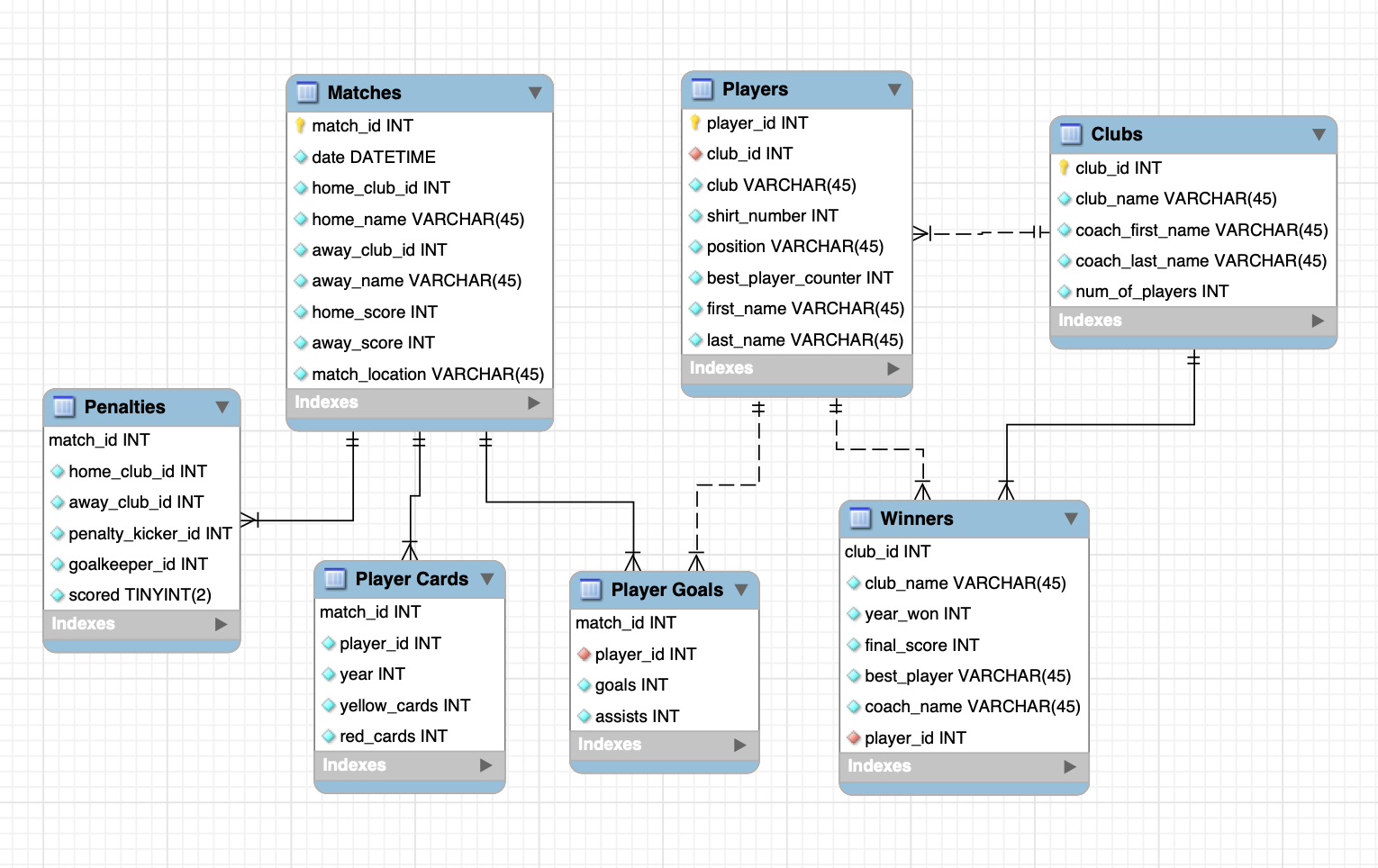
**Database Description**

****

Our team plans on incorporating data from Premier League football matches from 2015 to 2019 into our database. Our team’s objective with this database is to provide Premier League fans with a resource that allows them to obtain all important information regarding their favorite players or teams. Our database will include seven tables: Players, Player Goals, Clubs, Winners, Player Cards, Penalties and Matches.

The “Players” table will include the basic information about a player such as their first name , last name, current club , current shirt number, and position. This table will have a one to many relationship with four tables: Player Goals, Winners, Clubs, Penalties, and Player Cards using “player\_id”.

The “Player’s Goals “ table will include the number of goals and assist a player earned in a specific match. The relationship involved in this table will be a many to one relationship with the “Players” and “Matches” table using their respective primary keys.

The “Clubs” table will provide users with basic information of all the clubs in the premier league such as the clubs name, current coach’s name, and number of players. This table will have a one to many relationship with the “Winners”, and “Players” table using clubs\_id. This table will also have two separate one to many relationships with the ”Penalties” and “Matches” tables in order to represent the home and away clubs.

The “Winners” table will show the information of the club that won the league within a given season. This information will include the year the club won the league, the club’s final score in the league that season, and the name of the coach that won that year. The relationship within this table is a one to many relationship with the “Clubs” table using “club\_id”, and the players table using “best\_player\_id”.

The “Player Cards” table includes information about the number of red and yellow cards a player received during a specific match if they received a red or yellow card during that specific match. This table will have a one to many relationship with both the “Matches” and “Players” table.

The “Penalties” table will include information regarding the penalties that occurred during a specific match. This table will have a scored column with a boolean type which will record whether or not the player made the penalty kick as true or false. This table will have a one to many relationship with the “Matches” table. This table will also have two one to many relationships with the “Clubs” table to represent the home and away clubs, and the “Players” table to represent the penalty kicker and goal keeper.

The “Matches” table will include basic information regarding any match within the 2015 - 2019 season. This table will provide the users with the date of the match, home club, away club, the score of both teams and the location of the match. This table will have two separate one to many relationships with the “Clubs” table in order to represent the home and away club participating in the match.

**Sample Data Plan**

We still aim to use the sample data from Kaggle’s databases about the English Premier League. For each of the matches, we have data on the home teams, away teams, clubs, goals scored, winners, and results. There are plenty of databases for these specific columns that the creators on Kaggle found from the official Football Data website in the United Kingdom and the official English Premier League website. We were also able to find databases that had columns of coaches, players’ names and shirt numbers, location, and goalkeepers.

For columns that we don’t have enough data on, we plan to use other sources we research that aren’t just from databases and csv files. However, there are some columns that we found very little to no data on, which were the players’ cards for both yellow and red cards. Instead of taking out this table and its columns, we plan to look for other sources that will contain the data we need instead. We may also add in made-up data as well if absolutely necessary.

One of these sources we came across was the Premier League official website, which continues player stats and club stats. It also includes data for players cards, locations, goalkeepers, best players, and more that we couldn’t find on Kaggle’s website. There are rankings, filters, and other ways to organize their statistics which will be useful for utilizing it for our database.

**Progress Report**

Since our proposal submission, our team has furthered plans regarding the content and structuring of our database. We are continuing to refine our database conception, carefully considering possible necessary changes. Our team has also reviewed comments and feedback from our proposal review, and have met with our team advisor. One of the feedbacks we received from our peer review was that we should consider using additional sources to gather data. As such, we have considered incorporating other sources for information that is difficult to find alone through Kaggle.

Using our project proposal as a guide, we created an ERD of our database, which was an essential step in our process. We worked on this as a team, discussing the usefulness and pertinence of columns and tables. During our meeting with our advisor, we discussed concerns regarding possible redundancies in our data/ERD, and whether certain columns should be eliminated. Our advisor however recommended that our tables and columns are fine as is. Our team has communicated effectively using GroupMe throughout as well, in which we discussed meeting times, clarifications, etc. This has proven especially useful since everyone has busy schedules.

**Changes from the Initial Proposal**

There are a few minor changes that we made from our initial proposal. The first thing we did was normalize our tables and that caused some changes. There was a lot of redundancy throughout our tables that we had to fix. We have a table called “Players” that has information like the player id, their club id, first name, last name, shirt number, position and best player award count. We also had tables called “Players Goals” and “Players Cards” that also include the players id, first name and last name. To reduce the redundancy, we decided to take out the players first name and last name from both those tables since it is already included in the “Players” table. Just having the player id in those tables is enough information. This will make our database more user friendly and less crowded.

Another change we made from our initial proposal is how we were going to get our data. Our initial plan was to find datasets from places like Kaggle to fill our database. After lots of research, we realized that there was a lot of missing information out there. So far we found information that can help fill some of our tables, such as “Matches”, “Players”, “Clubs” and “Winners” tables. As of right now, our new plan is to find the missing information from other sources and manually plug it in the database or make up the missing data ourselves. We are going to definitely have to do this for our “Player Cards” table. Finding datasets for our database was a lot harder than we thought it was going to be but we are glad we figured it all out.

After these changes, we think our group is ready to start building the database using MySQL. Our tables now have a better flow and connect with one another better. Also, we now have a solution to how we are going to populate our entire database with information. We are excited to see what our finalized database is going to look like.

**Plans for Remaining Work**

Now that we have an idea of how our database will look like, our next step is to forward engineer our ERD to create our database and its tables. Once we have our database schema, we will start looking into the datasets we found of the Premier League and import them into our database using MySQL Workbench’s data import wizard. Before we attempt to import the data, we might need to do some cleaning up and modifications to the information kept in these datasets. We failed to find any datasets that fit our database perfectly, so we might need to reformat some of these datasets we found, perhaps change some data types and string formats so that it fits our database. Additionally, we might have to enter data manually after researching the relevant topic since some of the data we included in our ERD are not conveniently stored in a dataset we can find online, such as the number of yellow and red cards players received in a season. In our proposal we decided to look into the last 5 years of the Premier League, however if that proves to be challenging, we might reduce the scope of our project to include less data.  
 After we finally create our database and populate it with all the information we found online, we will need to start writing queries to answer some of the questions we asked in our proposal. To meet the requirements for the final report, these queries will include SELECT statements that are saved as views.

**References**

“Data Capture, Player &amp; Club Statistics: Premier League.” Data Capture, Player &amp; Club Statistics | Premier League, [www.premierleague.com/stats/clarification](http://www.premierleague.com/stats/clarification).

“Football Results, Statistics &amp; Soccer Betting Odds Data.” Football Betting - Football Results - Free Bets, [www.football-data.co.uk/data.php](http://www.football-data.co.uk/data.php).

Nalla, Zaeem. “Premier League Dataset.” Kaggle, 12 July 2018, [www.kaggle.com/zaeemnalla/premier-league](http://www.kaggle.com/zaeemnalla/premier-league).

Shubham, Maurya. “English Premier League Players Dataset, 2017/18.” Kaggle, 3 Aug. 2017, www.kaggle.com/mauryashubham/english-premier-league-players-dataset.